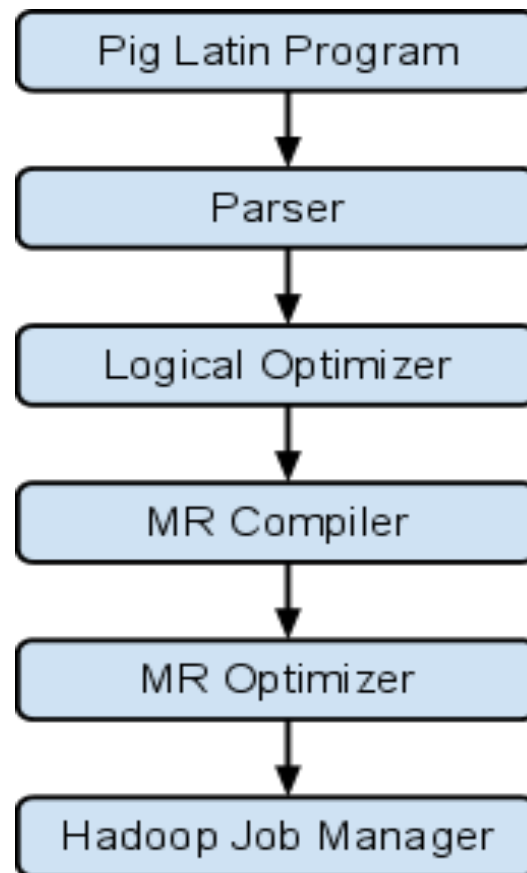


Introduction to Apache Pig



What is Pig

- A platform for analysing large datasets
 - A high-level language (Pig Latin)
 - An execution engine that generates MR jobs



Why Pig?

- Simple language
 - Easy for analysts familiar with SQL/Scripting languages
 - No need thinking in MapReduce terms
 - 10 lines of Pig ~ 200 lines of Java
- Support for rich, multivalued and nested operations
- Common Operations like JOIN, GROUP, FILTER, SORT are already provided.

An Example

Dataset: urls(url, category, pagerank)

Find the **top 10 urls by pagerank**, for each sufficiently large category:

```
urls = LOAD 'dataset' AS (url, category, pagerank);  
groups = GROUP urls by category;  
bigGroups = FILTER groups BY COUNT(urls)>10000000;  
result = FOREACH bigGroups GENERATE  
    group, top10(urls);  
STORE result INTO 'output';
```

An Example

Dataset: urls(url, category, pagerank)

Find the **top 10 urls by pagerank**, for each sufficiently large category:

```
urls = LOAD 'dataset' AS (url, category, pagerank);  
groups = GROUP urls by category;  
bigGroups = FILTER groups BY COUNT(urls)>100000000;  
result = FOREACH bigGroups GENERATE  
    group, top10(urls);  
STORE result INTO 'output';
```

Built-In statements

An Example

Dataset: urls(url, category, pagerank)

Find the **top 10 urls by pagerank**, for each sufficiently large category:

```
urls = LOAD 'dataset' AS (url, category, pagerank);  
groups = GROUP urls by category;  
bigGroups = FILTER groups BY COUNT(urls)>10000000;  
result = FOREACH bigGroups GENERATE  
    group, top10(urls);  
STORE result INTO 'output';
```

User Defined Functions

Pig Latin: A Not-So-Foreign Language

- High-level
- Declarative
- Consists of
 - **Statements** with Variables of specific **Types** and Built-In or User-Defined **Functions**

Pig Latin: Statements

- Loading and Storing
 - LOAD, STORE, DUMP
- Filtering
 - FILTER, DISTINCT, FOREACH...GENERATE, STREAM
- Grouping and Joining
 - JOIN, COGROUP, GROUP, CROSS
- Sorting
 - ORDER, LIMIT
- Combining and Splitting
 - UNION, SPLIT

Pig Latin: Types

- ☐ Numeric
 - int, long, float, double
- Text
 - chararray
- Binary ☐
 - bytearray
- Complex
 - tuple: sequence of fields of any type
 - bag: unordered set of tuples
 - map: set of key-value pairs

Pig Latin: Functions

- Eval

- AVG, CONCAT, COUNT, DIFF, MAX, MIN, SIZE, SUM, TOKENIZE

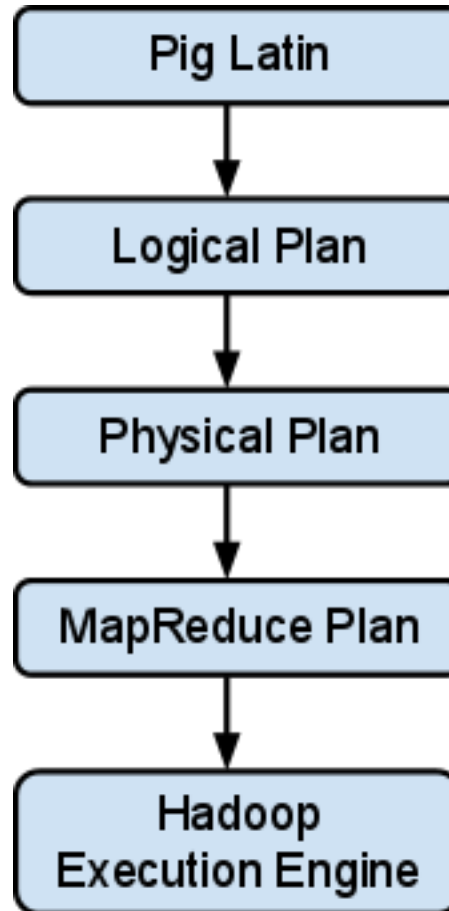
- Filter

- IsEmpty

- Load/Store

- PigStorage, BinStorage, BinaryStorage, TextLoader, PigDump

The Hadoop Compiler

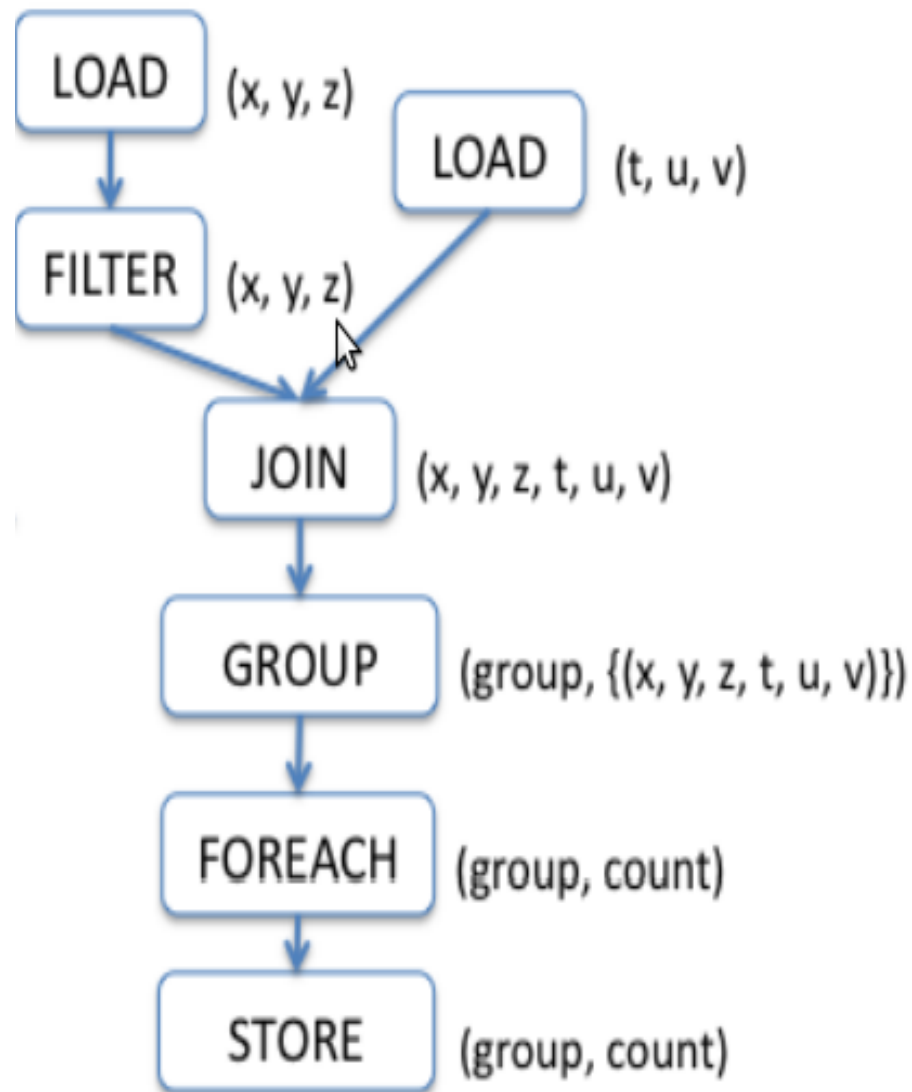


Pig Latin to Logical Plan

Pig Latin

```
A = LOAD 'file1' AS (x, y, z);  
B = LOAD 'file2' AS (t, u, v);  
C = FILTER A BY y>0;  
D = JOIN C BY x, B BY u;  
E = GROUP D BY z;  
F = FOREACH E  
  GENERATE          group, COUNT  
  (D);  
STORE F INTO 'output';
```

Logical Plan



Logical to Physical Plan

Logical Plan



Physical Plan

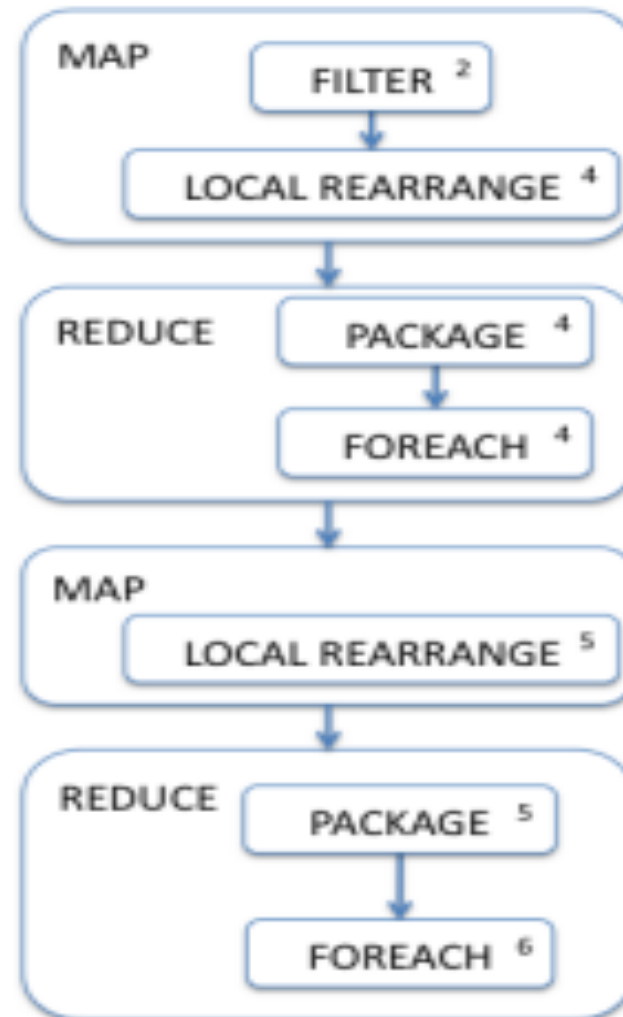


Physical to MapReduce Plan

Physical Plan



MapReduce Plan



MR Plan Compilation

- Convert each (CO)GROUP into a MapReduce job
- Map assigns keys based on the BY clause
- Each FILTER and FOREACH between the LOAD and the COGROUP are pushed into the map function
- Commands between COGROUP operations are pushed into the reduce function
- Perform tagging in case of multiple input sets
- Each ORDER command is compiled into 2 MapReduce jobs
 - Job 1 samples the input to determine key distribution
 - Job 2 generates roughly equal-sized partitions and sorts

Compilation Limitations

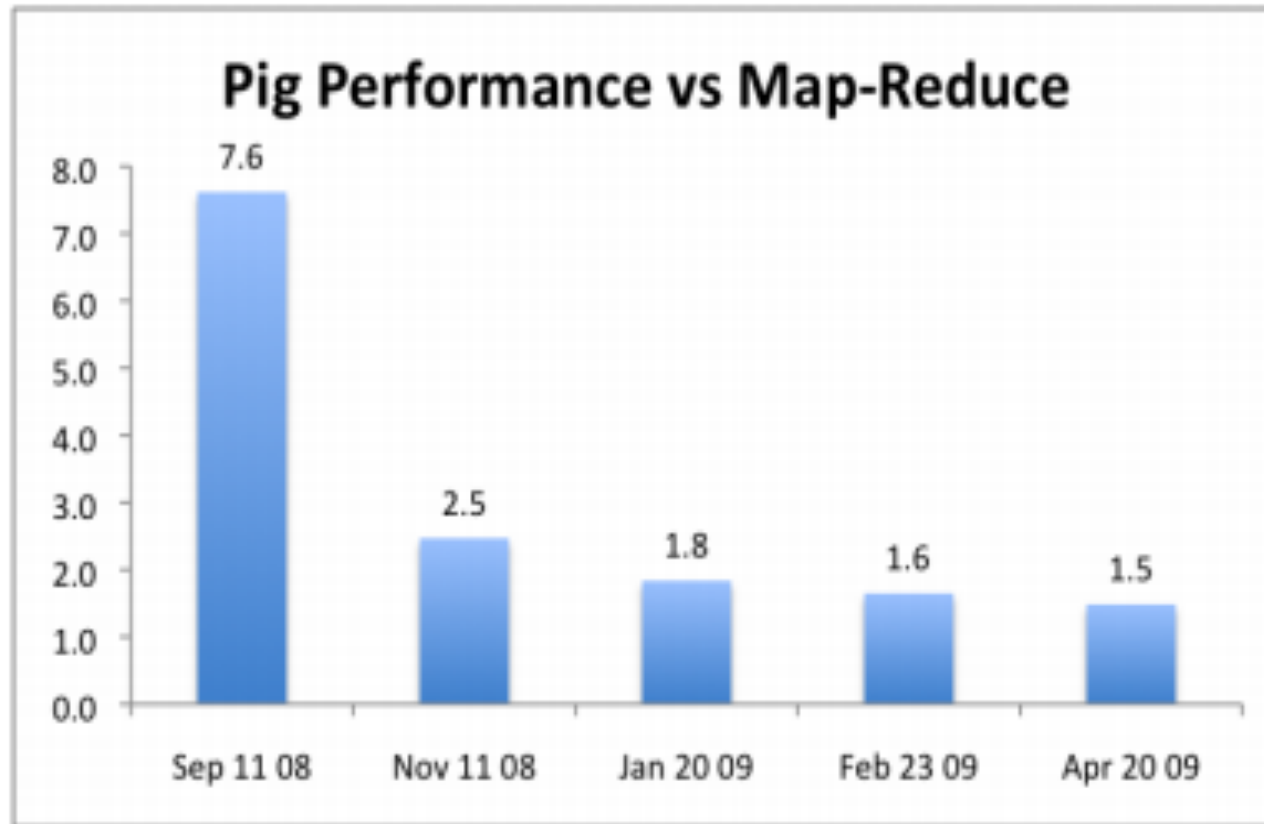
- Data must be **materialized** and **replicated** between MR jobs
- When dealing with multiple datasets, **tagging** is needed

Stratosphere might prove advantageous in these cases!

Evaluation - PigMix

- A set of queries to test Pig performance
 - latency
 - scalability
- Includes a set of MapReduce Java programs to run equivalent MapReduce jobs directly
- Used to measure the performance gap between running MapReduce directly and running Pig
- 4 datasets are provided with zipf or uniform distribution

Evaluation - Results



References

- Pig Latin Paper
- The Pig Experience Paper
- Hadoop, the Definitive Guide Book
- Cloudera Introduction to Pig video
- PigMix: <https://cwiki.apache.org/confluence/display/PIG/PigMix>